NATURAL LANGUAGE PROCESSING

BY- CHAITANYA MALHOTRA

DEFINITION

- *NLP*, is the sub-field of AI that is focused on enabling computers to understand and process human languages.

STEPS INVOLVED IN NLP

1. SEGMENTATION
2. TOKENIZATION
3. PREDICTION
4. LEMMATIZATION
5. IDENTIFYING STOP WORD
6. DEPENDENCY PARSING
7. NAMED ENTITY RECOGNITION
8. **COREFERENCE RESOLUTION**

SEGMENTATION

- The first step in the pipeline is to break the text apart into separate sentences.
- We can assume that each sentence in English is a separate thought or idea.
- It will be a lot easier to write a program to understand a single sentence than to understand a whole paragraph.

TOKENIZATION

- *London is the capital and most populous city of England and the United Kingdom."*
- The next step in our pipeline is to break this sentence into separate words or *tokens*. This is called *tokenization*. This is the result:
- *"London", "is", " the", "capital", "and", "most", "populous", "city", "of", "England", "and", "the", "United", "Kingdom", "."*
- Tokenization is easy to do in English. We'll just split apart words whenever there's a space between them. And we'll also treat punctuation marks as separate tokens since punctuation also has meaning.

PREDICTION OF PARTS OF SPEECH

- Next, we'll look at each token and try to guess its part of speech — whether it is a noun, a verb, an adjective and so on. Knowing the role of each word in the sentence will help us start to figure out what the sentence is talking about.
- We can do this by feeding each word (and some extra words around it for context) into a pre-trained part-of-speech classification model:
- The part-of-speech model was originally trained by feeding it millions of English sentences with each word's part of speech already tagged and having it learn to replicate that behavior.

LEMMATIZATION

- I had a **pony**.
- I had two **ponies**.
- Both sentences talk about the noun **pony**, but they are using different inflections. When working with text in a computer, it is helpful to know the base form of each word so that you know that both sentences are talking about the same concept. Otherwise the strings “pony” and “ponies” look like two totally different words to a computer.
- In NLP, we call finding this process *lemmatization* — figuring out the most basic form or *lemma* of each word in the sentence.

IDENTIFYING STOP WORD

- Next, we want to consider the importance of a each word in the sentence.
- English has a lot of filler words that appear very frequently like “and”, “the”, and “a”.
- When doing statistics on text, these words introduce a lot of noise since they appear way more frequently than other words.
- Some NLP pipelines will flag them as **stop words** —that is, words that you might want to filter out before doing any statistical analysis.

DEPENDENCY PARSING

- The next step is to figure out how all the words in our sentence relate to each other. This is called *dependency parsing*.
- The goal is to build a tree that assigns a single **parent** word to each word in the sentence. The root of the tree will be the main verb in the sentence.
- In addition to identifying the parent word of each word, we can also predict the type of relationship that exists between those two words.

NAMED ENTITY RECOGNITION

- *London is the capital and most populous city of England and the United Kingdom.*
- Some of these nouns present real things in the world. For example, "*London*", "*England*" and "*United Kingdom*" represent physical places on a map. It would be nice to be able to detect that! With that information, we could automatically extract a list of real-world places mentioned in a document using NLP.
- The goal of *Named Entity Recognition*, or *NER*, is to detect and label these nouns with the real-world concepts that they represent.

COREFERENCE RESOLUTION

- English is full of pronouns — words like *he*, *she*, and *it*.
- These are shortcuts that we use instead of writing out names over and over in each sentence.
- Humans can keep track of what these words represent based on context. But our NLP model doesn't know what pronouns mean because it only examines one sentence at a time.
- With coreference information combined with the parse tree and named entity information, we should be able to extract a lot of information out of this document.